Building a Manually Annotated Hungarian Coreference Corpus: Workflow and Tools

Noémi Vadász
Hungarian Research Centre for Linguistics
Hungary, Budapest
vadasz.noemi@nytud.hu

Abstract

This paper presents the complete workflow of building a manually annotated Hungarian corpus, KorKor, with particular reference to anaphora and coreference annotation. All linguistic annotation layers were corrected manually. The corpus is freely available in two formats. The paper gives insight into the process of setting up the workflow and the challenges that have arisen.

1 Introduction

The main motivation for building a coreference corpus was the fact that it is always interesting to investigate the behavior of a linguistic phenomenon in real texts. A manually annotated corpus is useful not only for linguists, but also for training and evaluating tools. KorKor, a Hungarian coreference corpus presented in this paper contains multiple linguistic annotation layers, such as disambiguated POS-tags, lemmata and morphological features (of two morphological tagsets) and dependency relations. All of these ordinary linguistic annotations were corrected manually, as well as the anaphora and coreference annotations.

Representativeness is an important feature of a corpus if we expect the tools trained on it to work with predictable quality in different genres and domains. However, in the current phase of the research, only two sources of texts were involved, since this phase aimed more at setting up the corpus building workflow and producing the necessary tools.

The resource is available under CC-BY 4.0 license to enhance accessibility, usability and extensibility. KorKor can be found in the following GitHub repository: https://github.com/vadno/korkor_pilot. Apart from the corpus itself, the whole workflow with detailed instructions, the annotation guidelines and the tools prepared in the frame of this project are also available in the GitHub repository to provide help for anyone having the necessary resources (financial resource, human labor, raw material) to continue the project or create a new, similar corpus based on it.

2 Background

2.1 Anaphora and Coreference

As a brief overview, here we discuss the definition of anaphora and coreference, which are often tangled in the literature. Resolution of both of them is required for interpreting a text, however the differences between them should be noted. An anaphora gets its interpretation from an other, previously mentioned constituent, its antecedent, therefore, it does not have an independent meaning. Coreference means that two expressions have the same referent. While anaphoric relations operate on the level of grammar, coreference belongs to the lexicon. As (van Deemter and Kibble, 1999) pointed out, coreference is a symmetric transitive relation, while anaphora is not, but it is context-dependent. An annotated corpus can contain e.g. only pronominal anaphora, but it can also be richly annotated with different relations between entities or even events. (Lapshinova-Koltunski et al., 2022) refers to the latter as "full coreference annotation", because it contains not only annotation of pronouns, but also full nominal phrases, verbal phrases and clauses and includes rich set of links with both entity and event coreference.

At the same time, in annotated corpora, occurrences in the text referring to the same entity are technically annotated similarly, and each type of anaphora is distinguished by different categories based on e.g. the type of the pronoun, as well as the different types of coreference relations. The differences between the two relation types are reflected in our annotation scheme in such a way that the type of the relation with the antecedent or previously mentioned coreferent element is marked next

Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2022) , pages 38–47
Gyeongju, Republic of Korea, October 16–17, 2022.

to the token. The labels used in KorKor for the different types of anaphora relations and coreference are detailed in Section 4.8.

2.2 Coreference Corpora

First, here we present the annotation schemes of two well-known shared tasks related to our topic. The annotation scheme of CoNLL-2012 (Pradhan et al., 2012) distinguishes between two types of coreference: Indentity and Appositive. The former is used for anaphoric coreference and all other types of mentions, the latter functions as attribution. The annotation scheme of MUC-6[1] and MUC-7 (Hirschman and Chinchor, 1998) does not separate different types of coreference. In these schemes coreference annotation is similar to a hyperlinked text, where the links connect the mentions of a given entity. An important objective of these shared tasks is to achieve high interannotator agreement, and following these schemes it can be accomplished. On the other hand, it is important to keep in mind that we have much more linguistic knowledge about the linguistic phenomena of coreference and anaphora, and these information can be important e.g. in information extraction tasks.

From the perspective of our work the most interesting resources are corpora of pro-drop languages, because the dropped elements as pronouns has referential properties. Hungarian is also a pro-drop language, which means that some pronouns (namely the personal and possessive pronouns in subject, object or possessor roles) can be left out from the sentence. In these cases, the person and number of the subject and the object can be calculated from the inflection of the finite verb, and the person and number of the possessor are calculable from the inflection of the possessum.

There are multiple coreference corpora for pro-drop languages, for example OntoNotes5.0 (Weischedel et al., 2013) for Arabic and Chinese, NAIST Text corpus (Iida et al., 2017) for Japanese, AnCora-CO (Recasens and Martí, 2010) for Spanish and Catalan, PCC (Ogrodniczuk et al., 2016) for Polish, and ParCorFull2.0 (Lapshinova-Koltunski et al., 2022) contains Portuguese as well. The annotation scheme of AnCora-CO includes dropped subjects in the syntactic trees and in the coreference annotation as well. NAIST, OntoNotes5.0 and PCC also contain zero pronouns.

ZAC (Zero Anaphora Corpus) (Baptista et al., 2016) is made specifically for the task of resolving dropped pronouns and contains texts in Brazilian Portuguese. It is 35,000 words long and contains texts from various sources. Only this linguistic phenomenon is annotated in it, however, it is really detailed, as it indicates the number and person of the dropped pronoun, indicates whether it is an anaphora or cataphora, and also indicates intersentential anaphoras separately, as well as providing the antecedent token. There are almost 1,500 zero anaphoras in the corpus, which clearly shows how important it is to deal with this phenomenon in the case of pro-drop languages.

As ParCorFull2.0 is a parallel corpus containing originally English and German texts, extending it with Portuguese was a challenge, because a pro-drop language had to fit into an annotation scheme which was not prepared to deal with this linguistic phenomenon. Here, the antecedents of the zero pronouns are marked next to the verbs, which seems to be a good solution, since the inflection of the verbs shows the characteristics of the dropped pronoun. This could only be applied to Hungarian by keeping in mind that a verb can have not only a dropped subject but also a dropped object, so it may happen that two antecedents need to be marked next to the verb.

It is also a possible solution, that the dropped pronouns do not appear in the corpus, since they are not present as independent tokens in the original text. This can be explained by the fact that the input of the coreference resolver does not contain dropped pronouns, and we do not necessarily want them to appear in the output, so we do not expect the resolver's training data to contain them either. On the other hand, for information extraction tasks, it is definitely useful if we have a richer linguistic annotation (e.g. zero verbs, ellipses and dropped pronouns). It can be a good solution that the corpus contains dropped pronouns but in a way that it can be used without them.

2.3 A Hungarian Coreference Corpus

The design of the corpus was inspired by the biggest Hungarian coreference corpus, SzegedKoref (Vincze et al., 2018). It was created by enriching a smaller part of Szeged Corpus (Csendes et al., 2005) with coreference

[1] https://cs.nyu.edu/~grishman/COtask21.book_1.html

annotation. It consists of student essays and newspaper articles giving altogether 55 763 tokens. 2 456 coreference chains were found in the texts, in which anaphoric and coreference relations are also included.

But why is another Hungarian coreference corpus needed besides SzegedKoref? Manually annotated data are always very valuable resources and the more of them, the better. Both SzegedKoref and KorKor have manually corrected annotation layers, therefore both of them are useful for numerous tasks apart from anaphora and coreference resolution. However, there are some differences between the annotation principles, schemes and tagsets, for instance in morphological and syntactic annotation. Joint use of the two corpora is still feasible after harmonizing the different formats.

Nonetheless, it has to be noted that there are some further differences between the two corpora on the level of theoretical issues. Both corpora contain dropped subjects, objects and possessors, but in contrast with SzegedKoref, in KorKor zero nodes for subjects are allocated to the infinitives, because they also play a role in the anaphoric relations. Another difference is that KorKor contains zero substantive verbs and ellipted verbs as well. Moreover, the method and the tagset of coreference and anaphora are different as well.

The tagset of SzegedKoref differentiates between the following relation classes: pronominal, nominal, adverbial, verbal and derivational. The class of nominal relations is divided into further subclasses: repetition, synonym, hypernym, holonym, epithet and apposition. In contrast, the tagset of KorKor contains only two tags for all nominal relations, which distinguishes identical reference and part-whole relation. However, the tagset of KorKor differentiates multiple types of pronominal anaphora with regard to the type of the pronoun: personal, demonstrative, reciprocal, reflexive and possessive, and it contains three extra tags for generic subject, speaker and addressee. The annotation guidelines of SzegedKoref highlights, that generic pronouns are not to be marked, but in our data we saw many examples that the generic subject in the text is also able to participate in anaphoric chains. Speaker and addressee is SzegedKoref got pronominal tag as other pronouns. Adverbial, verbal and derivational relations are not annotated in KorKor.

3 Data

3.1 Formats

The corpus is available in two formats. The setup of KorKor.`xtsv` follows the format used by the latest version of `e-magyar` (Indig et al., 2019), to be cited henceforward `emtsv`. In the `tsv` files, every line represents a token and sentences are separated by a blank line. Annotations are placed in the columns, which are described in the header. The motivation of using this format is that it fits well into the frame of `emtsv`, which was used during this project and which also can be used for further development of the corpus.

The KorKor.`conllup` files use the CoNLL-U Plus format[2]. A file of this format may contain any subset of the original columns of the core CoNLL-U files plus other project-specific ones. A comment listing the actual columns is inserted as the first line. This format is widely used, therefore the corpus could reach more people.

The two versions are different not only in their format but in their content as well, see the details in Section 4.9.

3.2 Sources

Texts from two sources were selected for building the corpus, using the collection of OPUS Corpus (Tiedemann, 2012): articles from Hungarian Wikipedia, and texts from the Hungarian website of the GlobalVoices[3] newsportal. Using OPUS ensures that the corpus is available under free licence. In addition to the coreference annotated corpus, a smaller amount of data (8,600 tokens) got only manually corrected lemmata, POS tags and dependency analysis. These data await further work, but at the same time the annotation layers completed so far could also be useful for others. Table 1. summarizes the size of the two formats of the coreference annotated corpus (in number of documents and tokens).

4 The Workflow

The building process was set up as a pipeline, in which as many steps were intended to be automated as possible. Human work was used for supervising and – if needed – correcting the annotation. Certain processing steps were carried out by the

[2] https://universaldependencies.org/ext-format.html

[3] https://hu.globalvoices.org

	documents	tokens (conllup)	tokens (xtsv)
huwiki	62	16,739	18,262
globv	32	7,760	8,799
TOTAL	94	24,499	26,581

Table 1: The size of the two formats of the coreference annotated corpus.

latest version of emtsv. As emtsv is a text processing pipeline, and the output of a given module forms the input of another one, it was reasonable to check and correct annotation not only at the end of the process but at several points of the workflow. Although human annotation in multiple cycles is certainly a labour-intensive method, minor faults are easier to fix, than muddled tangles. Thus, human annotators corrected the annotations in three phases.

The steps of the workflow were the following (tools used are in parentheses – steps where no tools are given were carried out with scripts developed within the project):

1. text collection
2. emtsv process (emToken, emMorph, and emTag modules)
3. format conversion
4. manual check (Google Spreadsheets)
5. format conversion
6. emtsv process (emDep module)
7. format conversion (emCoNLL module)
8. manual check (WebAnno)
9. manual insertion of zero substantives and ellipted verbs (plain text editor)
10. zero pronoun insertion (emZero module)
11. pronominal anaphora resolution
12. manual check and coreference annotation (Google Spreadsheets)
13. format conversion

The annotators have recorded the time needed for the correction of each document and each annotation layer. This information allows us to calculate the cost of the expansion of the corpus, and it could be helpful even in other corpus building projects.

annotation layer	token/hour
4. morphology	871.77
8. dependency	667.76
12. anaphora and coreference	595.86

Table 2: The time needed for manual correcting of the different annotation layers.

Table 2 shows the working hours needed to correct the different annotation layers.

The annotators reported every problem and question arising, therefore the annotation guidelines became finer and more detailed which sped up and made manual work easier.

The workflow includes multiple conversion steps between file formats, as the output of a certain step may differ from the expected input format of the following one. Each step of the workflow is specified below.

4.1 Preprocessing Texts

The selected texts consist of several sentences, because anaphora and coreference relations span through sentence boundaries. The length of the documents range from 5 to 27 sentences, the length of the sentences ranging from 3 to 71 tokens (counting punctuation marks as separate tokens). We paid special attention to add texts of manageable sizes to the corpus without truncation and wanted to include as many texts as possible from the sources. Therefore, in the case of both news and Wikipedia texts, we selected those that were of the appropriate length for our purposes, so we did not have to delete text fragments. Parts of some Wikipedia texts had to be cut out, but in these cases we made sure that the coherence and structure of the text did not change, and especially that there were no anaphoras without antecedents. The text selection was not influenced by the number of anaphora and coreference chains, as it was not checked in advance.

The texts were prepared for emtsv. Despite the fact that Wikipedia articles and news are edited texts, a lot of spelling errors had to be corrected

in them. Each text forms a raw corpus document (plain text files in UTF-8 character encoding).

4.2 Tokenization, Lemmatization and POS tagging

The output of the relevant modules of `emtsv` (emToken (Mittelholcz, 2017), emMorph (Novák, 2014; Novák et al., 2016; Novák, 2003) and emTag (Orosz and Novák, 2012, 2013)) is a `tsv` file of four columns (the format was described in Section 3.1). The content of the columns are: token, all possible lemmata and morphological tags, disambiguated lemma, disambiguated morphological tag.

4.3 Manual correction

In the first phase of manual work, tokenization, disambiguated lemmata and morphological tags were checked and corrected. Google Spreadsheets were used for this task, because it fits for most of our needs.

Seven linguists have edited the output of the modules of `emtsv` mentioned above. After some preprocessing steps that made the documents appropriate for Google Spreadsheets, conditional formatting was applied to make the document easier to follow and to give instant feedback to the annotators. Tokens for which the morphological analyzer produced multiple possible labels were highlighted. In case of tokens that have only one possible analysis anyway, the disambiguator is usually not wrong either. These tokens were not highlighted, but of course the annotators had to check them as well, since mistakes can occur in these too. Based on the annotators' feedback, conditional formatting and highlighting helped their work.

Besides tokenization, the disambiguated lemmata and morphological tags (the output of `emTag`) were checked by the annotators. To correct the lemma and the tag, they could choose from all possible lemma – morphological tag pairs of the token provided by `emMorph`. If none of them were acceptable, both of them could be set manually.

To make correction of tokenization errors easier, correcting commands were written into certain cells of the spreadsheet, e.g. to join or split tokens. First, the document was exported. Second, a postprocessing script responsible for the format conversion interpreted and carried out the correcting commands (such as line deletion, line insertion with the given content, joining two or more tokens, or splitting a token). The output format of the postprocessing script was again `xtsv`.

All the texts were corrected by at least two annotators, and a third one curated the documents. The inter-annotator agreement rate in terms of Cohen's κ for the morphological tags was: 96.07%.

4.4 Dependency Parsing

The corrected documents were fed into `emtsv` again for dependency parsing. As the dependency parser module (`emDep`) requires another morphological tagset in the input, the corrected tags were converted into a UD compatible tagset[4] by using a script[5]. Note that the UD tagset, in contrast with the `emMorph` tagset, does not encode derivational information, therefore the two layers differ not only in their format, but in their fineness as well. As the UD tagset is less detailed and lossless mapping is possible between them, no manual check was required. Thanks to this conversion step, end users can use two types of tagsets: emMorph, the current and most detailed Hungarian morphological tagset, and UD, which is widely used and meets an international standard.

4.5 Manual Correction and Zero Substantive Verbs

In this phase, WebAnno (Eckart de Castilho et al., 2016), a general purpose web-based annotation tool was used for manual correction, because it suited most of our needs. Link annotations as dependency edges are easy to handle with the drag-and-drop operation method, texts in different phases of analysis could be imported in various formats, and its interface allows us to check and correct already annotated documents as well. There are some additional functionalities like comparing and visualizing documents annotated by multiple annotators and calculating inter-annotator agreement. The flexibility of the tool provides that one can easily create a custom layer besides multiple built-in layers. WebAnno runs on a server and the annotators can use it via their common browser.

The output of the dependency module was converted to CoNLL-U, a file format edible for WebAnno. The conversion was done by the corresponding module of `emtsv`. Three linguists have checked and corrected the dependency edges.

[4]For details about Hungarian morphological tagsets, see (Vadász and Simon, 2019) and https://github.com/dlt-rilmta/panmorph.

[5]https://github.com/vadno/emmorph2ud2

Nevertheless, some weaknesses of the tool have turned out during the work. The tokenization was previously corrected, but still the annotators found tokenization errors in this phase as well. Unfortunately, WebAnno does not support token deletion or insertion, thus these errors had to be corrected in a separate postprocessing step.

In this postprocessing step, zero substantives and ellipted verbs were inserted as well. The reason why zero substantives and ellipted verbs were included is because they also have a subject – either overt or dropped – and ellipted verbs also have an object or other arguments.

A zero substantive was inserted in a sentence without a finite verb as a new token, where it would turn up as an overt substantive verb if the sentence was in past tense. Zero substantives got a combined ID from the ID of the preceding token. In Example 1, two zero substantives were inserted into the dependency tree.

Ellipted verbs are also insterted into the corpus, because in the absence of an overt verb, adjuncts could not be bound to their mother nodes. Ellipted verbs were also inserted manually, and as in the case of zero subordinates, they got a combined ID. In Example 2, an ellipted verb was inserted into the dependency tree.

Altogether, 419 zero substantives and 22 ellipted verbs were inserted into the corpus.

4.6 Inserting dropped pronouns

Dropped pronouns were inserted by a rule-based script[6]. The rules work on the preceding annotation layers (lemma, morphological tag and dependency analysis). Dropped pronouns are inserted in the following cases:

- subject, if a verb does not have a subject in the dependency tree;
- object, if a transitive verb does not have an object in the dependency tree;
- possessor, if a possessum does not have a possessor in the dependency tree;
- subject for an inflected or a non-inflected infinitive in the dependency tree.

Inserting dropped pronouns generates extra branches in the dependency tree. Zero subjects are placed after the verb, zero objects after the verb (and the subject), zero possessors after the possessum. All zero pronouns get a combined ID from the ID of the preceding token and the syntactic role of the zero element (SUBJ, OBJ, POSS). Not surprisingly, the POS tag of the zero pronouns is pronoun (PRON), their morphological features, like person and number, are calculated from the verb or the possessum.

Altogether, the corpus contains 867 zero subjects, 101 zero objects and 379 zero possessors.

4.7 Inserting pronominal anaphora

Pronominal anaphora relations are also inserted by a rule-based script. The script searches for the pronouns, and a set of rules operate on the POS tag, the morphological features and the syntactic information of the other words.

For the time being, the script searches for an antecedent only for personal pronouns, all other types of pronouns (possessive, reflexive, reciprocal, demonstrative and relative) had to be inserted manually. The antecedent searching algorithm for personal pronouns works by simple rules, e.g. if the subject of a verb is covert and the inflection of the verb is identical to the verb of the previous clause, the antecedent of the subject is the subject of the verb in the previous clause.

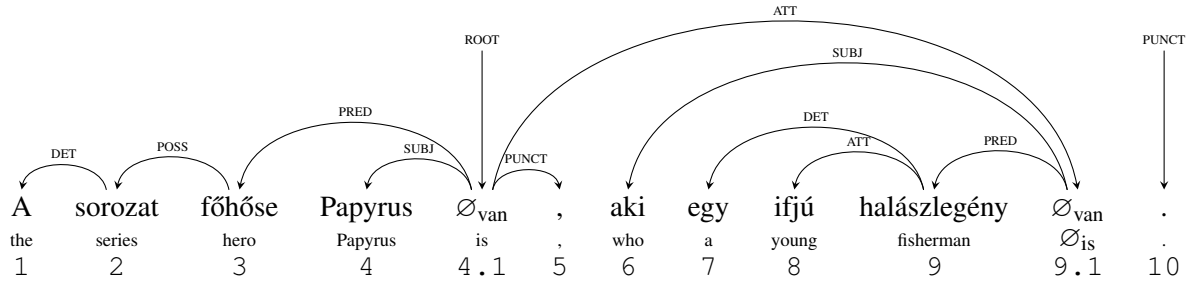
4.8 Manual Correction and Coreference Annotation

Four linguists have checked and corrected the insertion of the dropped pronouns and pronominal anaphora and annotated the coreference relations in this phase.

There is a large range of annotation tools capable for the task of anaphora and coreference annotation and some of them can be used not only for annotating but correcting already existing annotation as well. However, no annotation tools fit perfectly our needs, principally by reason of the inserted zero elements and the generated IDs.

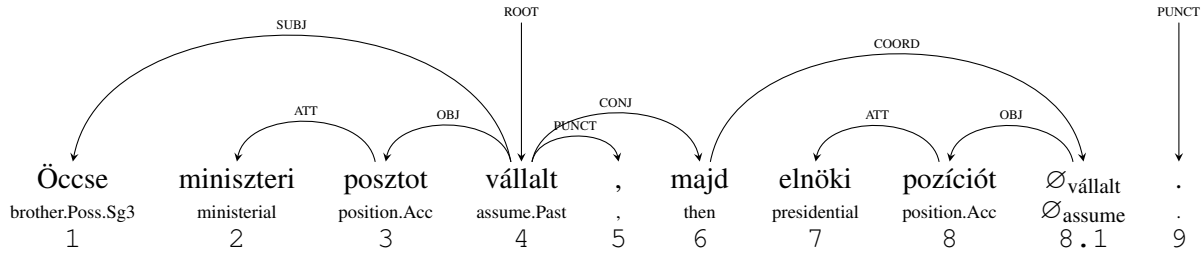
Hence, to perform this correction and annotation phase, Google Spreadsheets with conditional formatting was used again. Anaphora and coreference annotations were noted into two columns: one is for the ID of the head of the mother node, and one for the relation type. The following anaphora relation types are annotated in KorKor (with the tag in parentheses): personal (**prs**), demonstrative (**dem**), reciprocal (**recip**), reflexive (**refl**), relative (**rel**), possessive (**poss**).

[6]For the sake of anonymity, the link is provided only in the final version.



The hero of the series is Papyrus, who is a young fisherman.

Figure 1: In this complex sentence, the zero substantive verb of the subordinate clause is dependent from the zero substantive of the main clause. Original IDs and combined IDs of zero elements are under the tokens.



His brother undertaken a ministerial position, then a presidential one.

Figure 2: The verb of the first clause of the compound sentence occurs in the second clause covertly. A zero node is inserted, thus the arguments have a mother node to bind to.

The script that automatically inserted a link to the antecedent for the personal pronouns did not account for the other anaphora types and the relations in which they occur. For instance, the referent of a general subject – usually expressed in English by passive constructions – may be difficult to grasp. In Example 1 the verb *elítéltek* certainly has a third person plural subject, but it can not be related with any entities mentioned in the preceding text. In KorKor generic subjects are marked with the tag **arb**. General subjects do not have an antecedent, but they can be antecedents of other generic subjects.

(1) a Kínai Kommunista Párt egyik volt vezetője, akit hazaárulás miatt **elítéltek**
*one of the ex-leaders of the Communist Party of China, who was **convicted** for treason it was first **mentioned** in 1883 as an area donated to the Orthodox community*

Another interesting case is, when the speaker (or the writer) addresses the hearer (or the reader), as in Example 2. This type occurs rarely in the genre of news and Wikipedia, but still, some examples were found, moreover, expanding the corpus with other genres (literature, personal texts) would bring more instances.

(2) A születésnapi ajándékoknak is nagyon **örülünk**, ha **szeretnéd** támogatni a munkánkat, **küldj nekünk** adományt, vagy **vegyél** egyet az NSA-s karácsonyi **üdvözlőlapjaink** közül, amelyet a Creative Time-nál dolgozó **barátaink** terveztek.
***We**'re also very happy for birthday gifts, if **you** want to support **our** work, send us a donation, or buy one of our NSA Christmas cards designed by **our** friends at Creative Time.*

Two further tags were introduced to handle these special types of subjects: **addr** for the addressee, and **speak** for the speaker (writer). Bringing the addressee and the speaker/writer into the set of the participants of the event put down by the text allows us to mark if a pronoun refers back to these participants.

In coreference corpora, multiple types of coreference are usually annotated, such as repetition, varia-

tion, synonym, hypernym, hyponym, and holonym. While working out the design of KorKor and setting the annotation principles, we have faced some difficulties in connection with the different relation types, namely that it was challenging to write a guideline that could precisely define and differentiate the coreference types, because it is sometimes too hard for the annotators to distinguish the certain types. As a result, only two tags are used for marking coreference relations in KorKor. The tag **coref** is for the relation tpye when the two elements have identical reference (e.g. in the case of repetition, synonym, hiper- and hyponym). The tag **holo** is used when a part–whole connection holds between the two entities. It is important to distinguish these two types, because we found examples for "branching" coreference chains as in Example 3.

While in a coreference relation both participants are overt, the antecedent of a pronoun can be either a dropped pronoun or an overt phrase, therefore anaphoric and coreference relations make up a tangled net with branches, instead of a simple chain.

Table 3 summarizes the total number of each relation type in the corpus (counted in KorKor`.xtsv`).

relation type	occurrence
prs	1 306
dem	121
recip	10
refl	16
rel	294
poss	0
arb	274
speak	4
addr	1
coref	1 365
holo	180

Table 3: The total number of anaphoric and coreference relations in KorKor.

4.9 Converting to CoNLL-U Plus

The version of KorKor`.conllup` was converted from KorKor`.xtsv`. Although the two formats are interoperable, it was not only a simple format conversion. Firstly, zero elements are not listed as separate tokens in KorKor`.conllup`, which means that the affected dependency trees and anaphoric relations had to be revised and modified. Dropped pronouns are annotated in a different manner: if a verb has a covert subject or object, or if a possessum has a covert possessor, it is annotated in specific a column. Person and number of dropped subjects, objects and possessors are calculated from the inflection of the verb or the possessum. In the current state of the corpus these dropped pronouns are left out from the coreference chains, their antecedents are not marked and they can not be the antecedents of an other element.

Additionally, in KorKor`.conllup`, the coreferent elements form a simple chain, in which the elements having the same referent are linked linearly, instead of a tangled net structure with branches.

Consequently, the two versions fit for different users. KorKor`.xtsv` is suitable for examining the nature of anaphora from the linguistic point of view. The presence of zero elements allow the user to formulate queries about, for example, what events a participant in the text has attended. On the other hand, as KorKor`.conllup` is closer to to the usual coreference corpora, it is more applicable as a training or a test dataset, therefore is can form a base of a higher level information retrieval task, for example.

4.10 Further Questions

We made an interesting observation regarding Wikipedia articles, the annotation of which we often encountered serious difficulties. Illustrative example, when an article refers to an animal species, e.g. describes a certain type of chicken. First, it writes about the animal's features and habits in general, where it occurs, what it eats, etc., and then it covers the animal's body parts and their properties. The situation gets even more complicated if these are followed by presenting in detail separately the hen and the rooster (in first person sinular). These cases are marked as holonyms in KorKor, but this solution can be disputed.

Some problematic issues have emerged in connection with coreference, for which neither us, nor the literature have provided any answers yet. In Example 3, the state of the referent changes can be seen. What kind of relationship exists between a human and his/her dead body?

(3) Három hónap telt el **az újságíró házaspár**, Sagar Sarwar és felesége, Meherun Runi meggyilkolása óta. **A holttesteket** már exhumálták is, hogy megismételjék a boncolást.
*Three months have passed since the murder of **the journalist couple**, Sagar Sarwar and*

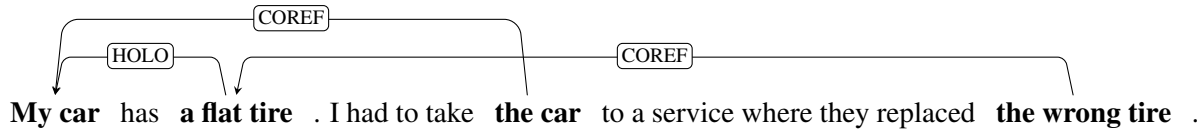


Figure 3: Branching coreference chains: a whole-part relation holds between *the car* and *the tire*, and both of them are repeated later in the text.

his wife. ***The bodies*** *are already exhumated to repeat the autopsy.*

Example 4 illustrates the issue of split antecedents.

(4) **Papyrus** bátor és megmenti **Thèti-Chèri-t**. **A két egymásra lelt barát** küldetést kap az istenektől, hogy védelmezzék meg a fáraót.
Papyrus *is brave and saves* ***Thèti-Chèri****.* ***The two friends found each other*** *got a mission from the gods to guard the pharaoh.*

According to our annotation principles, only one antecedent could be connected to a word, however the phrase *the two friends found each other* relates and refers to *Papyrus* and *Thèti-Chèri* at the same time. It would not help, if *Papyrus* and *Thèti-Chèri* were coordinated. In this case, the annotation would technically achievable, but it would be ambiguous, because the referring phrase could be either the whole coordination, or only the head of it.

Our annotation scheme does not cover the problem of these problematic cases, they are still waiting for solution and are part of our future plans, as is further expansion of the corpus.

References

J. Baptista, Simone Pereira, and Nuno J. Mamede. 2016. Zac : Zero anaphora corpus a corpus for zero anaphora resolution in portuguese. In *Proceedings of Workshop on Corpora and Tools for Processing Corpora, PROPOR 2016*.

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged Treebank. In *Proceedings of the 8th International Conference, TSD 2005*, pages 123–131, Karlovy Vary, Czech Republic. Springer.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan. The COLING 2016 Organizing Committee.

Lynette Hirschman and Nancy Chinchor. 1998. Appendix F: MUC-7 coreference task definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuji Matsumoto. 2017. Naist text corpus: Annotating predicate- argument and coreference relations in japanese. In *Handbook of Linguistic Annotation*, pages 1177–1196, Dordrecht. Springer Netherlands.

Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai. 2019. One format to rule them all – the `emtsv` pipeline for Hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Pedro Augusto Ferreira, Elina Lartaud, and Christian Hardmeier. 2022. Parcorfull2.0: A parallel corpus annotated with full coreference. In *Proceedings of the 13th Conference on Linguistic Resources and Evaluation (LREC)*, pages 805–813. European Language Resources Association (ELRA). Null ; Conference date: 20-06-2022 Through 25-06-2022.

Iván Mittelholcz. 2017. `emToken`: Unicode-képes tokenizáló magyar nyelvre (`emToken`: A unicode-compatible tokenizer for Hungarian). In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*, pages 70–78, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Attila Novák. 2003. Milyen a jó Humor? (What good humor is like?). In *I. Magyar Számítógépes Nyelvészeti Konferencia*, pages 138–144, Szeged. SZTE.

Attila Novák. 2014. A new form of Humor – Mapping constraint-based computational morphologies to a finite-state representation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Attila Novák, Borbála Siklósi, and Csaba Oravecz. 2016. A new integrated open-source morphological analyzer for Hungarian. In *Proceedings of the Tenth International Conference on Language Resources and*

Evaluation (LREC 2016), Paris, France. European Language Resources Association (ELRA).

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2016. Polish coreference corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 215–226, Cham. Springer International Publishing.

György Orosz and Attila Novák. 2012. PurePos 2.0 – an open source morphological disambiguator. In *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*, pages 53–63, Wroclaw.

György Orosz and Attila Novák. 2013. PurePos 2.0: a hybrid tool for morphological disambiguation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 539–545, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens and Antonia Martí. 2010. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language Resources and Evaluation*, 44:315–345.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Noémi Vadász and Eszter Simon. 2019. Konverterek magyar morfológiai címkekészletek között (Converters between Hungarian morphological tagsets). In *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)*, pages 99–112, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Kees van Deemter and Rodger Kibble. 1999. What is coreference, and what should coreference annotation be? In *Coreference and Its Applications*, pages 90–96.

Veronika Vincze, Klára Hegedűs, Alex Sliz-Nagy, and Richárd Farkas. 2018. SzegedKoref: A Hungarian coreference corpus. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0.